Int J Artif Intell Educ (2018) 28:1–28
DOI 10.1007/s40593-016-0121-0

ARTICLE

Designing Academic Writing Analytics for Civil Law Student Self-Assessment

Simon Knight[1] · **Simon Buckingham Shum**[1] · **Philippa Ryan**[2] · **Ágnes Sándor**[3] · **Xiaolong Wang**[1]

Published online: 3 November 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Research into the teaching and assessment of student writing shows that many students find academic writing a challenge to learn, with legal writing no exception. Improving the availability and quality of timely formative feedback is an important aim. However, the time-consuming nature of assessing writing makes it impractical for instructors to provide rapid, detailed feedback on hundreds of draft texts which might be improved prior to submission. This paper describes the design of a natural language processing (NLP) tool to provide such support. We report progress in the development of a web application called AWA (Academic Writing Analytics), which has been piloted in a Civil Law degree. We describe: the underlying NLP platform and the participatory design process through which the law academic and analytics team tested and refined an existing rhetorical parser for the discipline; the user interface design and evaluation process; and feedback from students, which was broadly positive, but also identifies important issues to address. We discuss how our approach is positioned in relation to concerns regarding automated essay grading, and ways in which AWA might provide more actionable feedback to students. We conclude by considering how this design process addresses the challenge of making explicit to

✉ Simon Knight
sjgknight@gmail.com

Philippa Ryan
first.lastname@uts.edu.au

Ágnes Sándor
agnes.sandor@xrce.xerox.com

[1] Connected Intelligence Centre, University of Technology Sydney, Broadway, Ultimo 2007 NSW, Australia

[2] Faculty of Law, University of Technology Sydney, Broadway, Ultimo, NSW 2007, Australia

[3] Xerox Research Centre Europe, 6 chemin Maupertuis, F-38240 Meylan, France

learners and educators the underlying mode of action in analytic devices such as our rhetorical parser, which we term algorithmic accountability.

Keywords Learning analytics · Writing analytics · Argumentation · Natural language processing · Rhetoric · Civil law · Participatory design

Introduction

Writing as a Key Disciplinary Skill

Critical, analytical writing is a key skill in learning, particularly in higher education contexts, and for employment in most knowledge-intensive professions (National Commission On Writing 2003; OECD and Statistics Canada 2010). Similarly in legal contexts, writing is both a 'tool of the trade', and a tool to think with – to engage in 'writing to learn' by considering the application of legal contexts through written legal documents (Parker 1997). A 1992 US report, commonly known as the MacCrate report (The Task Force on Law Schools and the Profession: Narrowing the Gap 1992), notes that although it is key for lawyers to learn effective communication methods (including analytical writing), there is in fact a disconnect between the practice, and legal-education of, lawyers with too little focus on this communication in legal training. The subsequent 'Carnegie report' (Sullivan et al. 2007) raised similar concerns, suggesting the need for reform in assessment practices with an increased focus on legal process and practice over product. Indeed, in the context described in this work, across the qualifications offered by the University of Technology Sydney (UTS) Law Faculty, critical analysis and evaluation, research skills (to find, synthesize and evaluate relevant information), and communication and collaboration (using English effectively to inform, analyse, report and persuade in an appropriate – often written – medium), are all highlighted as core graduate attributes. Thus, although there are stark differences internationally in the emphasis placed on writing in undergraduate legal education (Todd 2013), there are clear similarities between the English speaking common law countries and the emphasis on written communication in legal education. Learning the law is not simply about memorizing and recalling the contents of 'the law', but about thinking like a lawyer – the ability to process, analyse, and apply the law (Beazley 2004); abilities fundamentally tied to writing. Indeed, preliminary work indicates a relationship between grades in specific writing courses (common in the US context) and success in other law courses (Clark 2013).

Teaching academic writing is recognized as a challenge across higher education (Ganobcsik-Williams 2006) with a disparity between the more superficial presentational criteria by which students often judge their work, and the level of analytical argumentation that educators seek (Andrews 2009; Lea and Street 1998; Lillis and Turner 2001; Norton 1990). As a field, Law places huge emphasis on argumentation, but evidence suggests that its effective teaching has proven challenging. For example, a survey of US judges, practitioners and legal writing teachers indicated a universal generally poor view of new law graduates' writing skills (Kosse and ButleRitchie 2003). These respondents report writing that lacks: focus; a developed theme; structure; persuasive argument or analysis; synthesis and authority analysis; alongside errors in,

citation, grammar, spelling, and punctuation (Kosse and ButleRitchie 2003). Similar concerns are raised by other expert legal writers (Abner and Kierstad 2010).

A set of discussion and guidance literature has emerged for learning good practice in writing well in the law. These range from discussion of the *elegant* combination of clarity, concision, and engaging writing (Osbeck 2012), to very specific concerns regarding a preference for plain English over jargon and legalese (Stark 1984) – a concern shared by judges (across seniority and demography) who find plain English more persuasive (Flammer 2010). Others give specific guidance (see, for examples, Goldstein and Lieberman 2002; Murumba 1991; Samuelson 1984) which make clear that key elements of good legal writing include: Asserting a thesis (up front); developing an argument through use of analysis and synthesis of sources, facts, and legal argument (weighed in a measured way); and writing in a clear, simple, and direct or concrete tone.

To address concerns regarding written communication, legal-writing scholars have argued for an increased focus on the process of writing in both curricula and assessments. In the legal writing context (largely in American law schools) there have been calls for advice in writing mentoring to focus on underlying analysis, rather than structural features, (Gionfriddo et al. 2009); and for changes to assessment practices, with use of empirical studies to motivate (and assess the impact of) these changes (Curcio 2009); indeed, the same author has provided empirical evidence in the law-context that formative assessment can improve final grades by roughly half a grade (Curcio et al. 2008) with further preliminary evidence indicating a positive impact on mid-course grade (but not end of course) (Herring and Lynch 2014). Authors have thus suggested a need to address student's mindsets (Sperling and Shapcott 2012), and metacognitive and self-regulatory skills (Niedwiecki 2006, 2012) through effective formative assessment, with a commensurate desire to improve the level of self-reflection and professional writing development throughout one's legal career (Niedwiecki 2012; Vinson 2005).

Aligning Student and Instructor Assessments of Writing

At UTS students are usually admitted to a law degree on the strength of very good school-leaving results or upon successful completion of an undergraduate degree. As a general rule, both cohorts have strong writing skills. However, we identified that when students were invited to self-assess their own writing using the formal rubric they tended to over-rate their writing. If law students are not taught how to assess their own written work meaningfully while at university, they will be unlikely to learn this skill in practice. Yet it is in legal practice that the skill is most needed. The professional and ethical obligations that are imposed on legal practitioners mean that they must be mindful of what and how they write at all times. Most of what lawyers do involves reading, writing and critiquing correspondence, evidence, advice and instructions.

The metacognitive processes involved in assessing the quality of written work, particularly one's own, are sophisticated. Indeed, the scholarship on this point paints a negative impression of students' ability to improve their self-assessments. Research shows that people often have a faulty mental model of how they learn and remember, making them prone to both mis-assessing and mismanaging their own learning (Bjork et al. 2013). When students are taught to calibrate their self-reviews to instructor

defined assessment criteria, their learning outcomes improve (Boud et al. 2013, 2015). Importantly, self-review should be designed in such a way as to be formative in making critical judgments about the quality of the reviewed writing. A mechanism or intervention that causes students to pause and ask strategic questions about the content and quality of their writing could qualify as an incentive to proof-read and make the critical judgments required for meaningful self-monitoring. Ultimately, we seek to build students' ability to assess themselves as accurately as an expert assesses them, which as Boud has argued, is the kind of "sustainable assessment" capability needed for lifelong learning (Boud 2000).

One means by which to support such alignment is through the automated provision of formative feedback on the accuracy of students' self-assessment, or the writing itself. Indeed, a line of research has developed to analyse student writing through automated essay scoring or evaluation systems (AEE). These systems have been successfully deployed in summative assessment of constrained-task sets, with evidence indicating generally high levels of reliability between automated and instructor assessments (see, e.g., discussions throughout Shermis and Burstein 2013), with some criticism of this work emerging (Ericsson and Haswell 2006). Such systems have been targeted at both summative and formative ends. However, these approaches have tended to explore semantic content (i.e., the topics or themes being discussed), and syntactic structure (i.e., the surface level structures in the text), with some analysis of cohesion (see particularly, McNamara et al. 2014), but less focus on rhetorical structure (i.e., the expression of moves in an argumentative structure). Moreover, these systems have not typically been applied to formative self-assessment on open-ended writing assignments.

The Rhetorical Structure of Written Texts

The research described in this paper applies a natural language processing (NLP) tool for rhetorical parsing to the context of legal essay writing. The NLP capability in AWA is currently being developed as an adaptation of the rhetorical parsing module (Sándor 2007) of the Xerox Incremental Parser (XIP) (Aït-Mokhtar et al. 2002) to the legal domain. The parser is designed to detect sentences that reflect *salient rhetorical moves* in analytical texts (like research articles and reports).

The term *rhetorical move* was introduced by Swales (1990) to characterise the communicative functions present in scholarly argumentation. Swales defines rhetorical moves like *stating the relevant problem*, *showing the gaps* or *proposing solutions*. Rhetorical moves are usually conveyed by sequences of sentences, and often they are made explicit by more or less standardized discourse patterns, which contribute to the articulation of the author's argumentation strategy (e.g. *In this paper we describe …*- stating the relevant problem, *Contrary to previous ideas …* - stating the gaps, *In this paper we have shown …*- proposing solutions). The goal of the XIP rhetorical parser is the detection of the recurring patterns that indicate rhetorical moves in what we call *rhetorically salient sentences*.

Rhetorically salient sentences have successfully indicated relevant content elements in various text-mining tasks. For example, significantly new research is spotted by detecting a small number of "paradigm shifts" in tens of thousands of biomedical research abstracts (Lisacek et al. 2005) through the identification of salient sentences containing discourse patterns that convey contrast between past findings and new experimental evidence. Another application detects salient sentences that describe

research problems and summary statements. This application was tested for assisting academic peer reviewers in grasping the main points in research papers (Sándor and Vorndran 2009) and project evaluators in extracting key messages from grantees project reports (De Liddo et al. 2012). Moreover, as we describe later (Table 1) these moves may be mapped to a rubric structure in the legal writing context.

The analytical module of AWA[1] labels the following types of salient sentences (signalled in the text with highlighting and a 'Function Key – see next section): *Summarizing* issues (describing the article's plan, goals, and conclusions) (S), describing *Background* knowledge (B), *Contrasting* ideas (C), *Emphasizing* important ideas (E), mentioning *Novel* ideas (N), pointing out *Surprising* facts, results, etc. (S), describing an *open Question or insufficient knowledge* (Q), and recognizing research *Trends* (T). *Summarizing* is related to Swales' rhetorical moves *stating relevant problems* and *proposing solutions*, whereas all the other sentence types characterise *problem formulation,* which AWA's user interface refers to collectively as *Important Sentences*. Our typology of Important Sentences has been developed as a result of the detection of recurrent discourse patterns in peer reviewed research articles drawn from a variety of fields including social sciences and bio-medicine. Some examples of the discourse patterns are shown in Fig. 1.

The typology is robust in the text-mining tasks mentioned above (De Liddo et al. 2012; Lisacek et al. 2005; Sándor and Vorndran 2009) — but is designed to be modified if a new domain establishes the need for the detection of additional rhetorical moves. The rhetorical parser is the implementation of the concept-matching framework (Sándor et al. 2006), which models the salient discourse patterns as instantiations of syntactically related[2] words and expressions that convey constituent concepts. For example, sentences which contrasting ideas contain a pair of syntactically related words or expressions conveying the concepts of "contrast" and "idea/mental operation". Thus the following 3 syntactically and semantically different sentences are all labeled 'C′ by AWA, since the words in bold match this pattern: *challenge*, *need*, *failure* and *shift* convey "contrast" and *identify*, *highlights*, *demonstrating* and *notions* convey "idea/ mental operation". The two classes of words are syntactically related in all the three sentences:

C The second **challenge** is to **identify** different types of SLA and their associated technologies and uses.

C Consequently this **highlights** the essential **need** for repair.

C Finally **demonstrating** various solutions and the pragmatic **failure** or success of these with close regard to case law as well as the **notions** expressed by Keane in particular a **shift** of current ideology surrounding discovery.

These 3 sentences characterise analytical issues by identifying a challenge, highlighting a need, demonstrating failure and discussing the notion of a shift.

The question we investigate in this paper is whether it is possible to design automatically generated cues for civil law students and educators about the presence of valued qualities in

[1] AWA also has a module for analyzing reflective writing based on the Xerox Reflective Parser (Shum et al. 2016).

[2] Syntactic relationships are e.g. subject, object, modifier, preposition, etc.

Table 1 Mapping of assessment criteria rubrics to XIP salient sentence types and examples

Assessment rubrics: demonstrated qualities / standards	Associated salient sentence type	Examples (the discourse indicating the rhetorical moves is in bold)
Introduction		(S) (C) *Drawing upon the scholarship,* ***this paper*** *will* ***argue*** *that Australian court staff* ***should consider*** *using social media to increase confidence in the judiciary.*
• Statement of thesis	• *Summary* and *Important*	
• Essay plan	• *Summary*	
Content		(C) *However, the extent to which an intermediate appellate court may undertake to redefine the law as it sees appropriate – particularly when confronted with a judgment of a court of another jurisdiction but with equal standing in the judicial hierarchy –* ***raises*** *various* ***questions****.*
Development of sustained thesis	• *Important*	
• Knowledge, application and understanding of CPA, UCPR and common law	• -	(E) *Firstly, the* ***issue*** *is of general* ***importance****, and the fact that attempts are commonly made in corporate insolvencies to rely on this form of liability, makes a proper* ***understanding*** *of the second limb* ***important****, lest its application prove unjust.*
• Identification of relevant issues	• *Emphasis*	
• Critical analysis, evaluation and original insight	*Contrasting ideas*	(C) *Finally* ***demonstrating*** *various* ***solutions*** *and the pragmatic* ***failure*** *or success of these with close regard to case law as well as the* ***notions expressed*** *by Keane in particular a shift of current ideology surrounding discovery.*
• Development of coherent and persuasive argument	• -	
Conclusion		(S) ***In conclusion****, "law in practice" can be seen to have many differences and many similarities to "law in books" according to the snapshot I received from my visits to the Civil and Criminal Courts.*
Drawing together themes and answering the question posed in the introduction.	• *Summary* and *Important*	

student writing, and how these cues might serve as formative feedback to students when they are drafting their texts. In the remainder of this paper, we briefly introduce the AWA web application, describing its architecture and user interface. The evaluation of the tool is reported in terms of how we structured a legal writing academic's feedback to refine the rhetorical parser implemented in AWA, and the methodology for harvesting student feedback. We then analyse student feedback, before concluding with a discussion of how current limitations can be tackled, and the challenge of "algorithmic accountability", a broader concern in critical discourse about data in society.

The *AWA* Web Application

UTS has been developing an end-user application designed for student and staff use. The *Academic Writing Analytics (AWA)* tool is introduced below in terms of its NLP capabilities, architecture and user interface. AWA v1 described here is implemented in

Fig. 1 Examples of discourse indicators of rhetorical moves characterising research problems

PHP, while v2 is currently under development in Ruby-on-Rails. These operate in configurations approved by the university's IT Division to ensure that as we build confidence in its efficacy, it is ready for wider rollout as required.

AWA Architecture

AWA's architecture (Fig. 2) is designed to deliver the following capabilities, across all major web browsers and mobile devices:

- Authenticate users using university credentials, identifying their faculty
- Present discipline-specific sample texts users can experiment with, and discipline-specific examples in the user guide
- Accept plain text input from the user (pasted from source)

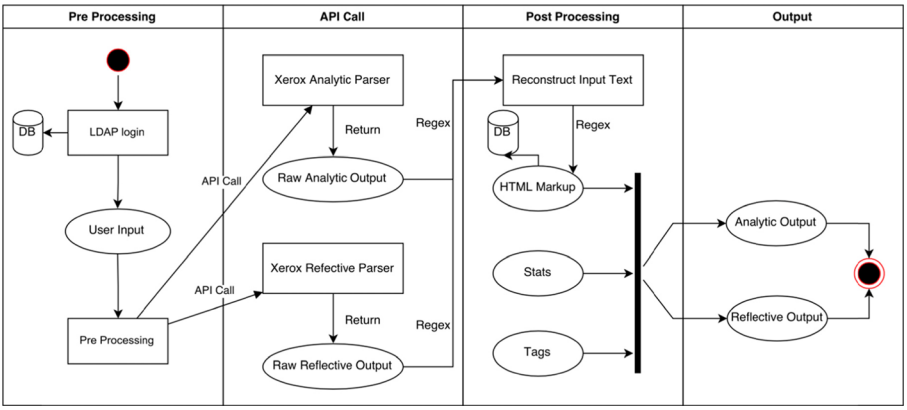

Fig. 2 AWA's functional architecture

- Log all submissions
- Invoke multiple NLP services on the Open Xerox server (to the reflective and analytic/rhetorical parsers)
- Render the output in multiple forms

The Rhetorical Parser

The rhetorical analysis in the Xerox Analytic Parser is implemented through the Xerox Incremental Parser (XIP) using syntactic parsing, dedicated lexicons and pattern-matching rules. Syntactic parsing extracts syntactic dependencies (such as the one between "challenge" and identify" in the sentence above), while the lexicons contain lists of words and expressions that are associated with the constituent concepts, and the pattern-matching rules select the sentences that contain dependencies that instantiate pairs of concepts necessary for conveying the labels assigned to rhetorically salient sentences (e.g. "contrast" + "idea/mental operation" = *Contrasting* idea). As described above, these rhetorical moves are: *Summarizing* issues (describing the article's plan, goals, and conclusions) (S), describing *Background* knowledge (B), *Contrasting* ideas (C), *Emphasizing* important ideas (E), mentioning *Novel* ideas (N), pointing out *Surprising* facts, results, etc. (S), describing an *open Question or insufficient knowledge* (Q), and recognizing research *Trends* (T). In the first prototype of AWA we have chosen to represent all the salient sentence types detected by XIP, however our analyses show that some of them are not particularly relevant in the legal domain. Thus in the future we might omit the irrelevant moves like (B), (Q) and (T), which are characteristic moves in empirical analyses, since their goal is the accumulation of knowledge, in contrast to legal analyses (and social sciences in general) where authors "negotiate" around facts and interpretations (Åström and Sándor 2009). The most frequent labels are those that are present in any kind of analytical writing, (S), (C) and (E).

AWA's User Interface Design Process

The NLP capability provided by the XIP rhetorical parser has been developed into a practical tool to produce a user experience good enough that students and academics are able and willing to engage with it. While the XIP rhetorical parser has been in development for over a decade, it is only over the last year that an end-user application for education has been developed.

In contrast to research prototypes and products for automated grading, we are designing AWA's feedback not as a summative grade, but as a means to provide formative guidance to improve draft writing. AWA adopts the metaphor of receiving feedback from a colleague who has used different coloured highlighters to draw attention to sentences she found interesting for some reason.

Although designed as a 'walk up and use' interface requiring no training, students are first introduced to AWA through a face-to-face briefing, sometimes in conjunction with instruction on academic writing that they would receive in any case from the writing support services in UTS. In this session, it is emphasized to them that it is a prototype in development, and that they should be critical of its output if they do not agree with it (see the discussion for further reflection on formative versus summative assessment).

On logging in for the first few occasions, students are welcomed as new users and prompted to visit the Guide which presents discipline specific sample texts and examples of each rhetorical move that AWA recognises. If the academics have not provided such examples they see default samples. The users paste in their text and submit it for analysis. AWA returns a range of visualizations, illustrated in Figs. 3 and 4. In addition, some basic statistics are visualized indicating rhetorical move frequencies, alongside (1) a wordcloud and (2) the key concepts, places, and people mentioned in the text.

The user interface design has been through many iterations, based on hours of testing with academics from many UTS faculties and units. A think-aloud method and screen recordings were used with teams as they worked alone, or in pairs/triads, to analyse and discuss sample student texts in AWA, while the researcher prompted them for thoughts, and explained how the system was working. We gradually debugged the design as we experimented with different ways to ensure that the users could navigate the system smoothly.[3]

Usability aside, the next question was whether AWA's output was experienced as academically trustworthy by the civil law lecturer, and her students. To date, we have reported statistical correlations between the frequency of certain XIP classifications and the quality of English literature essays (Simsek et al. 2015). However, user experience testing has not yet been reported; this application to the legal domain provides a first step to roll-out to students within a single domain.

Research Methods and Design

Participants and Research Design

In the research described in this paper, a collaboration between a law-faculty academic (the 3rd author), analytics researchers (the 1st and 2nd authors), a linguist (the 4th author) and an applications developer (the last author), we addressed the question of whether the AWA tool could usefully foreground the kind of rhetorical moves of interest in a legal assignment.

An alignment was first drawn between the assessment criteria, and the rhetorical moves identified by XIP, to establish the suitability of the tool for providing feedback on the particular task. The effectiveness of the tool was then evaluated with the law-faculty academic providing a close analysis of the accuracy of the parser for detecting the salient rhetorical structures in a sample of essays. Finally, a cohort of students was invited to engage with the tool, and provide their feedback, with 40 agreeing to do so and submitting their essays to AWA, as described in the sections on student evaluation and feedback.

Assessment Context

The law course in which the AWA tool is being developed has an established rubric, against which alignments with rhetorical moves were drawn. This rubric scores a number of facets on a 1–5 scale (unsatisfactory, satisfactory, good, very good,

[3] The system is presently available only for internal use. However, we are developing further resources (use guides, screencasts, etc.) which will be deployed on the public project website https://utscic.edu.au/tools/awa/.

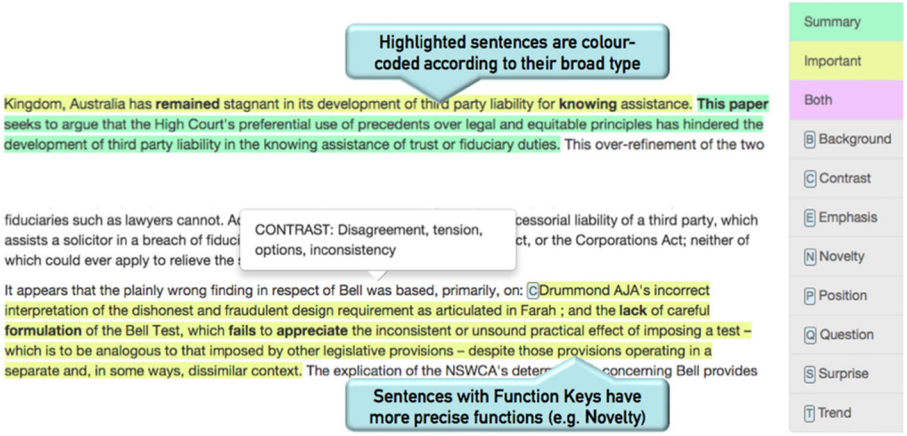

Fig. 3 A green *Summary* sentence signaling what the writer's intention is. On the right is the key to the different sentence types (clicking on this displays more details in the online Guide). Yellow sentences indicate the presence of a rhetorical 'key' (indicated below 'Both' in the key), for example the yellow. Pink sentences indicate that both a 'summary' and other important rhetorical move is made. *Contrast* sentence shown

excellent) aligned with the UTS grading scheme. The rubric is structured around particular kinds of move students should make in their texts, aligning these with sections that the students should use to structure their text. Essays were 2000 words in length, on a topic of relevance to civil-law, with a range of possible questions provided to students, for example:

> *The concept and meaning of good faith in negotiation and Alternative Dispute Resolution (ADR) processes, together with an articulation of what actions are required to comply with a good faith obligation or to support good faith negotiation, can be best described as an evolving 'work in progress' in Australia.*
>
> What actions do you think are required by practitioners in order to comply with these good faith obligations? Do you think it is possible for our courts to enforce

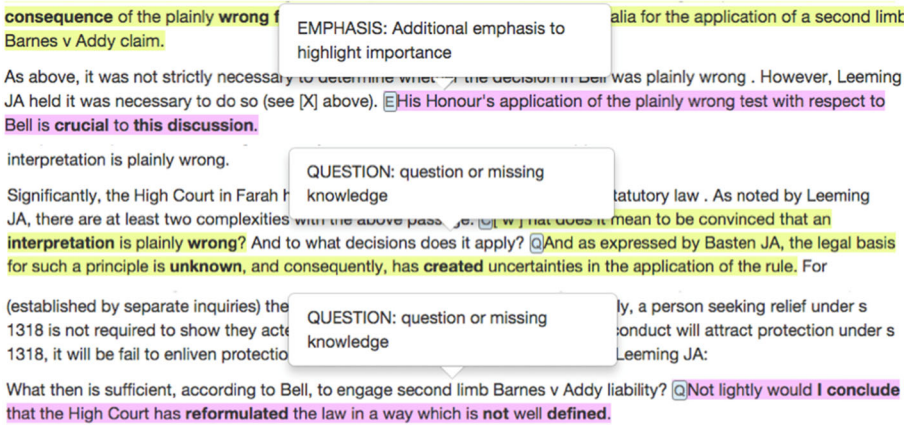

Fig. 4 Example sentences from civil law essays, classified by XIP and rendered in AWA

good faith obligations? Support your view with reference to this article and at least two additional recent authorities (For this purpose, "recent" means published in the last five years).

Students are thus asked to write an argumentative essay, forming a thesis or position in relation to the question. The rubric facets used in this course are:

1. INTRODUCTION: Statement of thesis and essay plan
2. CONTENT:
    a. Development of sustained thesis
    b. Knowledge and understanding of civil procedure act (CPA), uniform civil procedure rules (UCPR), and common law
    c. Identification of relevant issues
    d. Critical analysis, evaluation, and original insight
    e. Development of coherent and persuasive argument
3. CONCLUSION: Drawing together themes and answering the question posed in the introduction
4. REFERENCING: Written expression, footnotes and bibliography in accordance with Australian guide to legal citation (AGLC) 3rd edition

The relevance of the analytical rhetorical moves for legal essays is based on their association with the majority of the assessment criteria rubrics at UTS as shown in Table 1 which compares the elements of the writing rubric used in a civil-law assessment (column 1), with their associated salient sentence types (column 2), and gives an example instantiation (column 3).

Our observation from student self-assessment cover sheets indicates that students found self-assessment for these criteria challenging, since they overestimated their performance, and for the teachers, providing formative feedback on them may be prohibitively time-consuming. Effective (self)-assessment of legal writing requires the ability to recognise summary statements of introductions and conclusions, and the identification of text parts that contain critical analysis, and as a second step, the clarity and pertinence of the identified segments need to be evaluated. Both steps need expertise: the first mainly in the analysis of academic writing, and the second in domain knowledge. By highlighting sentences that need to be evaluated, AWA aims to provide support to the first step of this complex assessment activity, aligned with the guidance from the literature described in the introductory sections. Moreover, AWA also indicates in bold characters the relevant linguistic expressions that trigger the highlighting, with an aim to facilitate end-user understanding of the relevant parts of the highlighted sentences. The parser does not yet analyse or provide feedback above the sentence-level, as such it is left to students to reflect on whether sentences-types are positioned in the appropriate place at the whole-text, and section or paragraph level.

In the following sections we show how the salient sentence types noted above relate to structures inherent in any legal essay. We comment on some highlighted sentences of a sample legal essay from the LawTeacher.net web site.[4]

[4] http://www.lawteacher.net/free-law-essays/civil-law/law-in-action.php

Highlighting Statements of Introduction and Conclusion

A key feature of academic writing is conveyed, particularly in statements of introduction and conclusion, through widely taught rhetorical moves of academic writing such as "Outlining purposes", "Announcing present research", etc. (Swales et al. 2004). In the AWA tool, these moves fall under the *Summary* label. *Summary* sentences are expected in the introduction and the conclusion, as well as at the beginning and the end of sections.

The following *Summary* sentence is highlighted in the introduction of the sample student essay:

The **aim** of **this report** is to **assess** how "law in action" can be compared to "law in books" which may be done by observing the criminal and civil procedures of the Criminal and the Civil Courts.

The following *Summary* sentence appears in the conclusion of the same essay:

In conclusion, "law in practice" can be seen to have many differences and many similarities to "law in books" according to the snapshot I received from my visits to the Civil and Criminal Courts.

By highlighting these sentences AWA focuses the evaluator's attention on the rhetorical moves of introducing and concluding, while, as we have pointed out, it does not give any clue concerning the clarity of the statements. It is up to the evaluator to assess if these statements are relevant, well written, or if the conclusion matches the aims set in the introduction.

Highlighting Segments Containing Critical Analysis

Whereas the introduction and conclusion statements are clearly associated with acknowledged rhetorical moves of academic writing, and are identifiable in sentences, critical analysis is a more complex endeavour, which has not been associated with linguistic or structural units like sentences or specific section types. Thus pinpointing sentences that contribute to performing critical analysis is not straightforward. Critical analysis is usually explained in the form of general writing guidelines, like "indicate relevant issues", "formulate problems" or "present original insight". We suggest that sentences containing any of the salient rhetorical moves labeled in AWA except for the *Summary* move, are indicators of compliance with such guidelines: when the author points out relevant *Background* or *Contrast* between ideas, puts *Emphasis* on particular points, recognizes research *Trend*s and *Novelty*, she is indicating issues that she considers relevant; when she describes *Contrast*s and hitherto unanswered *Question*s, she is formulating research problems; or with *Contrast* and *Emphasis* she introduces original insight. We do not claim that our list of rhetorical moves indicating particular aspects of critical analysis is exhaustive, since it is the distillation of corpus studies in previous text-mining work. Should a new aspect emerge in corpora, it could be integrated into the framework.

The following examples from the sample essay illustrate how such moves reflect aspects of critical analysis in the sample essays. The sentence below is labeled *Emphasis* and *Contrast*. It introduces the discussion of relevant issues in what follows,

and points out the importance of discussing some other issues. Although the "relevant issues" themselves are not contained in the highlighted sentence, the sentence still indicates that the author does handle them as an integral part of the essay, and thus the reader can identify and evaluate them in the subsequent sentences. This sentence also draws the reader's attention to the necessity of discussing an analytical aspect ("the differences between the two jurisdictions"), which is another indicator of the treatment of relevant issues in the analysis:

E C Before **discussing** the **relevant issues** in the Criminal and Civil courts, it **is necessary** to **discuss** the differences between the two jurisdictions which will enable us to discover why the procedures and processes differ between them.

The following sentence is labeled *Contrast*, and it formulates a problem (that of "judicial independence"), which signals that the author is engaged in critical analysis:

C This **questions** the **issue** of judicial independence as to whether or not judges reach decisions in an independent way, only taking into account the facts and the law **rather** than their own opinions or the opinions of government, political parties, businesses, organizations and the media.

We emphasize again that the highlighted sentences convey elements of critical analysis based on the rhetorical moves they contain, and the assessment of the relevance of these elements with respect to the topic of the essay remains an expert task.

Evaluation Methodology with The Law Academic

Confusion Matrix Annotation

We have developed a practical methodology for refining the quality of the parser, using a form of semantic annotation by the domain expert (the civil law academic leading the course) of AWA's output. Designed to be practical for the development of analytics tools with staff with limited time and resource, this is a rapid version of the more systematic coding that a qualitative data analyst would normally perform on a large corpus, using signal detection theory codes for True/False Positives/Negatives, to populate a confusion matrix:

		Law Lecturer	
		Important	Unimportant
AWA	Selected	TP	FP
	Unselected	FN	TN

Thus, the lecturer was asked to highlight *True Positives* and *Negatives* where she agreed with AWA's highlighting and classification of a sentence, or its absence; *False Positives* where AWA highlighted a sentence she did not consider to be significant, and *False Negatives* where AWA ignored a sentence she considered to be important. We placed misclassifications of a sentence in this class too, as another form of failure to spot an important move.

We did not prepare particular annotation guides for the lecturer, since we cannot provide very detailed explanations of AWA highlights for the students or teachers either. As we described above AWA labels are based on complex patterns which would be far too cumbersome to describe in AWA. Our aim is to keep the AWA labels intuitively understandable, which is a challenging aspect of the UI. So, we defined the rhetorical moves informally in one sentence and gave some examples for each type. This first experiment served also as a test if the label names, the short explanations and the examples in AWA enable an analyst to grasp the semantics of the labels. We wanted to gain insight into the ways to improve the guide in the UI (rather than formally assessing the annotation scheme).

Starting with the generic rhetorical parser, the lecturer selected several essays with known grades. She pasted AWA's output into Microsoft Word, and using the agreed scheme, annotated it with TP/FP/TN/FN plus explanatory margin comments. The linguist in turn commented on this document, for instance, explaining why AWA behaved the way it did when this confused the lecturer, or asking why a sentence was annotated as FP/FN.

This structured approach to analyst-academic communication began to build a corpus from which one could in principle calculate metrics such as precision, recall and F1; however, it is not yet large enough to calculate these reliably. Rather, the confusion matrix provided more focused feedback than informal comments to the team, aiding rapid iteration and dialogue, using a familiar tool (Word) and a simple 2 × 2 representation that required no training. We return to the importance of this process in the discussion on algorithmic accountability.

Refinements to AWA

For each of the cells in the confusion matrix, we consider examples of successes and failures, and the different adaptation measures that were taken to improve the signal/noise ratio.

True Positives

Consistent with the intentions of the rhetorical analysis these sentences illustrate correct classification:

Contrasting ideas:

C **However**, the extent to which an intermediate appellate court may undertake to redefine the law as it sees appropriate – particularly when confronted with a judgment of a court of another jurisdiction but with equal standing in the judicial hierarchy – **raises** various **questions**.

Emphasis:

In section II, **this essay** will **outline** the fundamental characteristics of mediation and the role of a mediator.

Summing up the main topic of the essay:

CE **Discovery** involves an exchange of a list of **documents** between the parties to a case, which are **relevant** to the **issues** in **dispute**.

False Positives

We found that sentences annotated as False Positives by the lecturer were falsely triggered by patterns that are often relevant in non-legal academic writing, but in law the same patterns are used as legal 'terms of art', for instance:

CE **Discovery** involves an exchange of a list of **documents** between the parties to a case, which are **relevant** to the **issues** in **dispute**.

We can reduce False Positives in such cases by deleting the legal terms from the XIP lexicon, but the complication is that these words may also be used in their analytical sense. In such cases we implement disambiguation rules. In the following sentence "issue" is not used as a legal term, and so the sentence should be (and is) highlighted:

CE These **fundamental problems** frequently **surface** the legal **landscape**, as even fairly small disputes can **raise issues requiring** the examination of numerous electronic resources to identify valid documentation.

False Negatives

A few false negatives were due to the fact that analytical content in legal essays may use different words or expressions for conveying the constituent concepts from those that are parts of the existing lexicons. For example, neither "assess" nor "argument" was part of the lexicon, and thus the following sentence was not labeled. Once the words are added, the SUMMARY pattern is detected by the parser, and the sentence is labeled.

Section V **assesses arguments** in favour and against judicial mediation and deliberates the compatibility of roles.

While one aspect of adaptation is the expansion of the lexicon, in fact the overwhelming majority of false negatives were due to sentences that the law academic coded as relevant in terms of her interpretation of the XIP categories, but which do not contain patterns coded in XIP.

For example, the lecturer expected the following sentence to be labeled as 'C′:

Whilst technology has facilitated the retention of all records for businesses, Keane firmly **maintains** its' **converse** effect.

This sentence does indeed convey a contrast. However, it is not labeled, because the contrast is not between two "ideas", but between one effect of technology (i.e. it "has facilitated the retention of all records for businesses ") and Keane's maintaining a "converse effect" of technology. Technically speaking even if this sentence does contain words that represent the relevant analytical concepts, it is not selected, since

there is no syntactic relationship between any two of them. We can consider that this sentence partially fulfils the criteria for being selected, since it contains words instantiating some constituent concepts.

Were the sentence formulated in more precise terms, i.e. as a contrast between "ideas", it would be highlighted, and labeled as 'Contrast', thus:

C **Whilst** it is generally **considered** that technology has facilitated the retention of all records for businesses, Keane **maintains** its converse effect.

In this case we need to consider the extension of the current analysis, because it seems that the AWA patterns are too restrictive for the 'C′ move.

The following sentence was expected by the lecturer to be labeled as 'B′ *Background knowledge*:

Discovery involves an exchange of a list of documents between the parties to a case, which are relevant to the issues in dispute.

This general description of the concept of "discovery" can legitimately be interpreted as "background knowledge", however, it does not have the same semantics as 'B′ in AWA. The semantics of the 'B′ label in AWA is "reference to previous work", as illustrated in the true positive 'B′ sentence:

B Previous studies have shown that the phonological deficits that characterise dyslexia persist into adulthood.

The role of the sentences annotated as false negatives in legal essay analytics needs to be elucidated before further adaptation is undertaken. On the one hand we need to revise the UI definitions and explanations so that they are in line with the users' intuitions, and on the other hand, we need to consider the modification of discourse patterns to be detected in order to target more specifically legal discourse.

Taken together, the existing general analytical parser without any adaptation did provide relevant output for legal tests. Our data are too small for computing meaningful metrics, thus in Table 2 we report the result of the annotation exercise in terms of numbers of sentences.

This test indicated that lexical adaptation is required: deleting legal 'terms of art' from the general lexicon, and extending the lexicon with genre-specific vocabulary used in legal essays for conveying rhetorical moves. No syntactic parse errors were the cause of any

Table 2 Result of the sentence annotation exercise

		Law Lecturer	
		Important	Unimportant
AWA	Selected	TP: 19	FP: 13
	Unselected	FN: 7	TN: 52

False Negatives or False Positives. Even if some sentences in the legal essays are longer than average general texts sentences, this did not have an effect on the parser performance.

We started the lexical adaptation based on the test annotations. We created shared documents where we collected words to be added and deleted as we encountered them during the development process. Table 3 illustrates the list of legal 'terms of art' collected for deletion.

Currently, the implementation of changes (such as those introduced above) to XIP is performed by hand. We foresee the elaboration of mechanisms that automatically update the lexicons on user input or learn the domain vocabulary through machine learning.

No formal evaluation of the effect of the changes has been performed, but it is interesting to analyse the output of the slightly adapted parser on the document used for the annotation exercise. Having updated the lexicons following some basic adaptation the confusion matrix showed the results indicated in Table 4.

These changes resulted in a decrease in the number of False Positives with a commensurate increase in the number of True Negatives. This was due to the deletion of the legal terms from the general analytical lexicon. For example, the following sentence was highlighted as 'Contrast' in the general analytical parser, but not in the adapted legal parser, because of the elimination of *issue, solution* and *problem* from the lexicon.

It will further consider the benefits and constraints of the early identification of issues in court proceedings, and pre - hearing conferences as potential solutions to these problems.

The remaining False Positives and False Negatives are due to the differences of the definition of the rhetorical moves between the annotator and the general analytical parser. Further analysis is needed to determine if it is necessary to change the scope of the analytical parser by adapting the patterns to legal rhetorical moves.

Having taken the first steps in refining AWA for the legal domain, we moved into a first iteration of exposure to the students themselves.

Evaluation Methodology with The Students

Introducing AWA to Students

The evaluation of AWA by Law students was designed carefully to ensure that it did not disadvantage any students. Students had already been introduced to the concept of text analysis software in a legal technology lecture, setting the context for an open invitation to engage critically with a prototype. They were informed that they would only be given access to AWA *after* submission of their essays, to avoid any perceived risk of

Table 3 – List of 'legal terms of art' to be deleted from the general analytical lexicon

Claim, conduct, contest, contribution, discover, discovery, dismiss, dispute, dispute resolution, document, documentation, evidence, issue, issues in dispute, limits the term, limit(ation), method, page limit, problem, quest, represent(ation), resolution, resolution of dispute, resolve, role, solution, solve, term

Table 4 Change of the result of the annotation exercise after some lexical adaptation in terms of the number of sentences

		Law Lecturer	
		Important	Unimportant
AWA	Selected	TP: 19	FP: 5
	Unselected	FN: 7	TN: 60

unfair advantage. AWA was thus framed not only as a tool for self-assessment of their own writing, but as an example of the emerging tools which they could expect to encounter in their professional practice, particularly those who choose careers in commercial litigation.

Forty students volunteered to participate in the project (submitting essays to AWA) and of those initial volunteers, twenty managed to attend introductory sessions where they were introduced to the impact that NLP is having on jobs in diverse sectors, in education specifically, and then introduced to AWA and shown how to use it, concluding with open discussion. Both sessions were held after the participants had submitted their essays, verified against student records.

In the sessions it was emphasized, on the one hand, that AWA was a prototype and students should not be unduly concerned if it failed to highlight sentences they believed to be sound; on the other hand, the academic law staff present indicated that they had growing confidence in it based on their testing.

Survey

The students were given a week to experiment with AWA, and were then sent a 4-question online survey, with 12 of the 40 participants submitting responses:

1. How comfortable are you with getting feedback of this sort from a computer?
2. Did you find the feedback meaningful, so you could see ways to improve your writing?
3. If we continue to make AWA available, what is the likelihood that you would use it?
4. We'd love to hear any further thoughts you have. Please comment here.

Reflection Statements

In addition, all students on the course were invited but not required to orally present 2 min 'reflection statements', worth 5 % of the total subject grade. AWA pilot students could choose to reflect on their experience of AWA, as an alternative to reflecting upon other material in the course which other students did. Reflective statements were assessable based on oral presentation only (no written version required), all assessed against the same criteria: use of plain English expression, speaking style, description of content upon which the student was reflecting and clear statement of what is understood as a result of engaging with that learning content (or the use of the AWA tool). Students were also invited to state how their understanding of legal practice might be

influenced by their learning or experience. Of the 280 students taking Civil Law, 277 students chose to complete a reflection statement, with 8 of the 40 AWA-students choosing to specifically reflect on their experience with AWA, 2 of whom also provided written copies of their statements, while data from another 5 comes from the lecturer's notes. All students gave written permission for their feedback to be used anonymously for research purposes.

Qualitative Coding of Student Feedback

An analysis was conducted of the survey data (including comments), the written student reflections, and the lecturer notes from the oral presentations on AWA. For those completing the questionnaire, response frequencies were tabulated. Further analysis of the written content was conducted in NVivo (QSR International Pty Ltd 2012) to identify thematic patterns across the content; these are reported in the next section, with broad patterns noted and exemplifications of the feedback given in brackets.

Student Feedback: Results

We organise the feedback data into several themes we discerned:

- AWA's value as a tool for provoking reflection
- AWA's lack of sensitivity to linguistic nuance
- Student sentiment on receiving this kind of feedback from a machine
- Student appetite for automated support
- Student uncertainty on the relationship between AWA output and final grade

AWA's Value as a Reflective Tool

Survey question: *Did you find the feedback meaningful, so you could see ways to improve your writing?*

N of students	1	7	3
Rating	Not all meaningful	Meaningful in part	Yes, it was all meaningful

A number of students mentioned in their written comments or reflections that the AWA feedback had helped them to think differently about their writing by using the highlighting – and lack of highlighting – as a prompt for reflection. Table 5 illustrates the students' views on the value of this.

Lack of Sensitivity to Linguistic Nuance

Interestingly, students mentioned that they had reflected on their writing, even though they questioned the accuracy of the AWA feedback ("I found it really useful in

Table 5 Student feedback on the value of AWA highlighting rhetorical moves

"I found it interesting to note that the AWA tool picked up problems with my essay that I had not noticed." *Student 5 reflection notes*
"I definitely found it useful. It also made me realise that I tend to use bold, certain language in making my point towards the end of each paragraph rather than up front at the beginning (when introducing my point)." *Respondent 5*
"I also tend to signpost a point with Important language and then actually make the point, rather than just making the point (by that I mean, the sentence highlighted as Important was often the one just before the sentence I would have thought was making the important point, before using this tool)." *Respondent 5*
"I realise now what descriptive writing is - the software had quite a bit to say about my lack of justification - also true - pressed for time and difficult circumstances have caused this for me in this instance - good to see it sampled." *Respondent 9*
"[I] wanted to make my essay slightly different from a formal research essay because of the content being about innovation - i adopted a partial prose / commentary style with loads of referencing. I see that the style of my argument was weak but i feel i still made the point that I wanted to. With it as a tool I could tighten my argument and research style especially if it becomes more informed about what difference in styles may mean and then be able to critique or offer feedback in different ways." *Respondent 9*
"It was very easy to use and understand how to work the program. It seems like a great way for students to visually reflect on their writing and be able to straight away see where they should of (sic) included more critical analysis and language of emphasis." *Respondent 10*
"I felt that the feedback was meaningful as it highlighted how formal my academic language for the purposes of my assignment was. The feedback showed the places in my essay where I was critical and also helped me realise that I should of (sic) been more critical in other areas of the essay. The blank parts of the essay that were not highlighted by the program also showed that I may have needed to be more critical in those parts." *Respondent 10*
"The tag cloud tab in the program also was good in that it showed me how broad my vocabulary was in the essay and whether I needed to build on it." *Respondent 10*
Said feedback was instructive…" because of the way the information is presented by breaking down the sentences and clearly marking those that are salient as being contrast or position etc" *Respondent 11*
"I put through different types of academic papers I have written and discovered that I did not use recognised summarising language consistently across different faculties." *Respondent 12*
"I like the idea of this program and can see that it is useful at identifying some elements of writing" *respondent 12*

scrutinizing my own work and culling the fat from my essay. I don't think it was 100% accurate in what it did, and the bold words gave me a really good indication of the sort of phrasing it was looking for" *Respondent 5*).

Although another student (who had marked that the AWA feedback was 'not all meaningful') noted that AWA was "not able to identify a single introductory remark" *respondent 7*, while both they and an expert colleague had thought the writing contained such features. Another student (who marked the feedback as 'meaningful in part') noted:

> "it is possible to make a clearly stated point in an academic way without using one of the markers (and it is possible that tools such as this have not been programmed to search for all possible permutations of metadiscourse) that would be recognised by the algorithm. I think perhaps that saying that if a paper does not use specified 'signposts' suggests that the writing is not clear and 'academic' (see 'tips' on the results page), constricts writing style. I think it is possible to be clear about your position without explicitly saying 'my position is…'". *respondent 11*

Other students made similar claims:

> "…found that the tool was limited in its ability to pick up on summary sentences. It was able to detect phrases such as 'ultimately, this essay will conclude,' or 'in conclusion,' but the use of adverbs such as 'thus,' and 'evidently,' in conclusive statements failed to be recognized."…"Another limitation is that certain sentences, which were recounts or mere descriptions were deemed important, whilst more substantive parts of the essay containing arguments and original voice failed to be detected." *Student 1 reflection*).

On Receiving Feedback from a Machine

Survey question: *How comfortable are you with getting feedback of this sort from a computer?*

N of students	0	1	1	5	2
Rating	Not at all comfortable	Not really comfortable	Indifferent	Quite comfortable	Very comfortable

Some students were very positive about removing the personal element ("takes the emotion out of having your work scrutinized" *respondent 12;* "it was not embarrassing in the way that it can be when a tutor or marker gives feedback" *student 7 reflection notes*); and the potential for on demand feedback ("feedback is available 24 hours a day" *student 7 reflection notes*; "I think AWA will eventually be able to help bridge the 'teaching/learning' divide [between large classes & few tutor consultation hours]" *student 4 reflection notes*). Some students also noted the reflective possibilities of using AWA in an ongoing writing process, for example:

> "…writing multiple drafts and proof reading each can be both tiresome and difficult considering it is often hard to recognize the flaws of your own writing when you've been working on it for so long. Xerox's tool acts as a great, objective third party in providing early feedback." *Student 1 reflection*

> "I would be comfortable receiving feedback of this sort from this kind of tool in the same way I'm comfortable receiving feedback from a person - it is something to reflect on and consider in order to make decisions whether implementing the suggestions/feedback will improve your piece of writing, or your writing generally." *Respondent 11*

One noted the potential of AWA to fit into their current writing workflow, noting "I currently run most of my essays through 'Grammarly' [a grammar-checking website] before submission" *respondent 6.* However, some students provided the caveat that they would consider AWA as one source of feedback alongside others, and/or that they would not "trust it as I would trust feedback from an academic" *respondent 12*.

Appetite for Automated Support

Survey question: *If we continue to make AWA available, what is the likelihood that you would use it?*

N of students	0	0	2	4	6
Rating	Not likely at all	Not very likely	Not sure	Quite likely	Very likely

There was a clear appetite for support from tools such as AWA among these students. The students were invited to use the tool to reflect on their current assignment; however, a number of them mention 'testing the system' – uploading multiple assignments; varying expression in individual assignments; and uploading assignments from varying disciplines, to explore the differences in feedback obtained. Indeed respondent 8 (noted above with regard to a desire to connect feedback to outcomes), said:

> "I dug out a few older essays from different law subjects and ran them through the software. Some where I got average marks (60-70%) and another where I absolutely excelled and got 95%. When I compared these essays, I didn't see much difference in the stats analysed by the software – all my work seemed to have quite low detection rates of 'importance', yet on some I got 60%, while others 95%." *Respondent 8*

Another reported that "I put through different types of academic papers I have written and discovered that I did not use recognised summarising language consistently across different faculties." *Respondent 12.* Indeed, one student looked up the research papers AWA is based on (listed on the AWA site), saying:

> "Overall I'm impressed with the tool and think it would be a powerful instrument for students, markers and academics. Originally it appeared to me to essentially be searching for words, but after looking more carefully at the results I can see that it is analysing sentence structure to provide commentary, which is impressive" *respondent 11.*

Relationship to Grade

Perhaps not surprisingly, students wanted to know how the AWA feedback related to outcome ("I would only find real value in this software if it improved my grades. If framing my writing with 'contrast', 'emphasis', 'position', etc gave me better marks then the feedback would be very meaningful." r*espondent 8*; "I would like to know if the changes I would have made would have improve my mark." *Student 8 reflection notes*).

Limitations of the Evaluation

The student evaluation was conducted in an authentic context, with students reflecting on how an assignment that they had just submitted might have been

improved had AWA been available. This has generated extremely useful insights, but we recognise that this is a preliminary evaluation with a small sample size. While in other case studies we have been able to test for statistical patterns and calculate classification metrics, the annotated corpus from this work is not yet large enough to do this reliably. We thus have qualitative evidence of AWA's value from the law academic and students, which has yet to be quantified. The emerging themes indicate potential areas for targeting future evaluation, providing qualitative insight into the potential and areas for improvement in the AWA tool. We now consider the implications of the student feedback, and reflect on the state of this project in relation to broader concerns about whether machine intelligence such as this can be trusted.

Discussion

The prospect of automated analysis of writing finding a place in mainstream educational experience provokes strong reactions and natural scepticism. In writing we can express the highest forms of human reasoning: how can a machine make sense of this in a meaningful manner?

The work presented here has sought to describe a user-centered design process for a tool that is not seeking to grade, thus removing some of the 'high stakes' controversy surrounding automated grading (see, for example, Condon 2013; Deane 2013; Ericsson and Haswell 2006). However, seeking 'only' to provoke reflection on the part of the student writer in no way removes the obligation to ensure as far as possible that this is productive reflection about meaningful prompts: if the signal-to-noise ratio is too low, students and educators will find they are wasting their time reviewing meaningless highlighted text, and will disengage. The student feedback indicates that AWA was provoking useful reflection, but was not without its problems. AWA can in no sense be considered complete, and we are acutely aware of its current limitations, which set the agenda for our ongoing work.

From Highlighting to Actionable Reports

A key challenge that we are now considering is how to bridge the gap between the current ability to highlight sentences, and capability to generate a meaningful report which is more clearly actionable by the writer. A number of approaches to this are under consideration. At a simple level, 'canned text' may be deployed, triggered by the recognition of a simple pattern (e.g. the absence of any *Summary* sentences in the abstract or conclusion). Advancing our current analysis, combining sentence-types for analysis at the paragraph or multi-sentence level may prove fruitful. In addition, more advanced Natural Language Generation approaches would permit greater flexibility in the conditional construction of feedback to students (e.g. failure to use *Contrasting* rhetorical moves when discussing particular authors or theories known to be in tension with each other — from prior expert knowledge of the field). Progress on this front will help to address uncertainty from students and instructors as to how to make sense of the highlighting.

"Does this Highlighting Mean it's Good?"

Related to the previous point, but standing as a question in its own right is the extent to which students and educators should be encouraged to use rhetorically-based highlighting as proxies for the overall quality of the piece. Prior work (Simsek et al. 2015) has investigated statistical relationships between the frequency of all or particular XIP sentence types, and essay grade, with some positive correlations found, but clearly there is much more to the creation of a coherent piece of writing than just this indicator, so one does not expect it to account for all variance. Rhetorical parsing on its own does not assess the truth or internal consistency of statements (for which fact-checking or domain-specific ontology-based annotation (Cohen and Hersh 2005) could be used). Other writing analytics approaches provide complementary lenses (see, for example, McNamara et al. 2014) which, when combined in a future suite of writing analytics tools, would illuminate different levels and properties of a text in a coherent user experience.

We are considering deploying automated techniques to analyse the patterns of highlighting in XIP output. For example, sequential analysis might detect patterns in the sequences in which different highlights co-occur, or follow each other. We can also hypothesize that certain sentence types may occur in particular kinds of writing, or particular sections of a given genre of document.

Addressing Algorithmic Accountability

Algorithms sit at the heart of all analytics, but their design and debugging remains the preserve of the few who possess statistical, mathematical or coding expertise. In an era when data collection pervades societal life, embedded in appliances and the physical environment, it is impossible to understand how all the algorithms 'touching' our lives work. Some might ask if learners or educators should be troubling themselves with why software is behaving as it does, if it is behaving predictably. However, when their functioning becomes a matter of enquiry or concern, these cannot remain black boxes. For many, there is an ethical need to articulate the possible definitions and meanings of "algorithmic accountability" (see, for example, Barocas et al. 2013; Diakopoulos 2014). For learning analytics, this is also a pedagogical need, such that learning analytics have appropriate levels of transparency to different stakeholders.

In the context of AWA, we propose three ways to respond to this challenge, noting also their limitations.

Open Research Publications The way in which XIP operates has been published in the research literature (Aït-Mokhtar et al. 2002), as well as the contextualization to academia in this case study. While this is a dissemination channel suited for researchers, and citeable peer reviewed research adds credibility for non-academics, AWA's function requires translation into appropriate forms for educational practitioners and students who also have the right to enquire. Currently this takes the form of the website's Guide, and personal briefings presented to academics and students as orientation.

Openly Testable System Behaviour Many of XIP's services are publicly accessible via a public API (Xerox n.d.), providing another form of accountability: it can be tested

at will by anybody able to use the API. The rhetorical parser documented here is available only to research partners at present, but rigorously testable. XIP is not, however, available in open source form at present, which is unquestionably the most rigorous form of accountability for suitably qualified people to inspect.

Open Stakeholder Communication Most academics and students do not benefit from being given source code, just as they do not benefit from being given SQL backups. Transparency is a function of ability to benefit from the information, and accountability a function of the quality of response to queries, and ultimately the consequences of failing to give an adequate account, or of causing harm of some sort. Thus, users build confidence and ultimately come to trust software because they trust the way in which it has been developed, and the tool adds value to their teaching/learning. The *academic's* trust in AWA has grown through extensive discussion with the *learning analytics research team*, experimenting with early AWA prototypes, receiving satisfactory answers to questions, and seeing that her feedback is acted on at the levels of both the user interface and behaviour of the parser. We have also described how we used the structured annotation scheme to scaffold communication between the *academic* and *linguist*. AWA is thus experienced as accountable because as questions arise, they are answered and/or code is modified.

The *linguist* tuning XIP is another stakeholder: her trust in the process is similarly built as her algorithms are tested in authentic contexts, and enthusiastic end-users are giving detailed feedback to improve its performance. We have completed only one design iteration with the *students*, but we anticipate that engaging them in future iterations will build their confidence in a similar manner.

Most software tools using a participatory, agile approach go through this kind of design process in their early stages. The implications for analytics products whose designers are not available to answer questions or requests from educators and students remain to be seen. Many companies are now putting in place the human and technical infrastructure to remain responsive to user communities of thousands, challenging as that is. As we have discussed elsewhere (Buckingham Shum 2015), it may be that many educators and students do not in fact want to examine the algorithmic engines powering their tools, as long as they seem to be working — or it may be that we must find new ways to make the black boxes transparent in ways that satisfy the curiosity and literacies of different audiences.

Conclusion

To conclude, in the context of civil law undergraduate writing, we have documented the design process through which we are developing and evaluating AWA, a writing analytics tool intended to provide formative feedback to students on their draft writing. In order to reach the point where we could confidently pilot this with students, we evolved a participatory design process using structured annotation of AWA's output by the law academic, which we believe could find wider application as a practical technique. The piloting of AWA with students provided valuable feedback for future improvements, and parallel AWA extensions to other disciplines, which are now in development.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Abner, E., & Kierstad, S. (2010). Preliminary exploration of the elements of expert performance in legal writing, A. *Legal Writing: Journal of Legal Writing Inst, 16*(1), 363–411 Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jlwriins16&section=13.

Aït-Mokhtar, S., Chanod, J.-P., & Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering, 8*(2–3), 121–144 Retrieved from http://journals.cambridge.org/abstract_S1351324902002887.

Andrews, R. (2009). *Argumentation in higher education: improving practice through theory and research.* New York: Routledge.

Åström, F., & Sándor, Á. (2009). Models of scholarly communication and citation analysis. In *ISSI 2009: The 12th International Conference of the International Society for Scientometrics and Informetrics* (Vol. 1, pp. 10–21). BIREME/PAHO/WHO & Federal University of Rio de Janeiro. Retrieved from http://lup.lub.lu.se/record/1459018/file/1883080.pdf

Barocas, S., Hood, S., & Ziewitz, M. (2013). Governing algorithms: a provocation piece. Available at SSRN 2245322. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2245322

Beazley, M. B. (2004). Better writing, better thinking: using legal writing pedagogy in the casebook classroom (without grading papers). *Legal Writing: Journal of Legal Writing Inst., 10*, 23 .Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jlwriins10&section=9

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-113011-143823.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151–167. doi:10.1080/713695728.

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education*, *38*(8), 941–956. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/02602938.2013.769198.

Boud, D., Lawson, R., & Thompson, D. G. (2015). The calibration of student judgement through self-assessment: disruptive effects of assessment patterns. *Higher Education Research and Development*, *34*(1), 45–59. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/07294360.2014.934328.

Buckingham Shum, S. (2015). Learning analytics: on silver bullets and white rabbits. Medium. Retrieved from http://bit.ly/medium20150209.

Clark, J. L. (2013). Grades matter; legal writing grades matter most. *Miss CL Rev, 32*(3), 375–418 Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/miscollr32&section=24.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6*(1), 57–71. doi:10.1093/bib/6.1.57.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: fishing for red herrings? *Assessing Writing, 18*(1), 100–108 Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293512000505.

Curcio, A. A. (2009). Assessing differently and using empirical studies to see if it makes a difference: can law schools do it better? *Quinnipiac Law Review, 27*(4), 899–934 Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1487778.

Curcio, A. A., Jones, G. T., & Washington, T. (2008). Does practice make perfect? An empirical examination of the impact of practice essays on essay exam performance. *Florida State University Law Review, 35*(2), 271–314 Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1135351.

De Liddo, A., Sándor, Á., & Buckingham Shum, S. (2012). Contested collective intelligence: rationale, technologies, and a human-machine annotation study. *Computer Supported Cooperative Work (CSCW), 21*(4–5), 417–448. doi:10.1007/s10606-011-9155-x.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24. doi:10.1016/j.asw.2012.10.002.

Diakopoulos, N. (2014). Algorithmic accountability. *Digital Journalism, 3*(3), 398–415. doi:10.1080/21670811.2014.976411.

Ericsson, P. F., & Haswell, R. H. (2006). *Machine scoring of student essays: truth and consequences*. Utah State University Press .Retrieved from http://digitalcommons.usu.edu/usupress_pubs/139/?utm_source=digitalcommons.usu.edu%2Fusupress_pubs%2F139&utm_medium=PDF&utm_campaign=PDFCoverPages.

Flammer, S. (2010). Persuading judges: an empirical analysis of writing style, persuasion, and the use of plain English. *Legal Writing: Jorunal of Legal Writing Inst*, *16*(1), 183–222. Retrieved from http://www.law2.byu.edu/Law_Library/jlwi/archives/2010/183.pdf.

Ganobcsik-Williams, L. (Ed.) (2006). *Teaching academic writing in UK higher education*. Basingstoke, UK: Palgrave Macmillan.

Gionfriddo, J. K., Barnett, D. L., & Blum, J. (2009). A methodology for mentoring writing in law practice: using textual clues to provide effective and efficient feedback. *Quinnipiac Law Review, 27*(1), 171–226 Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1264390.

Goldstein, T., & Lieberman, J. K. (2002). *The Lawyer's guide to writing well*. University of California Press .Retrieved from http://www.ucpress.edu/excerpt.php?isbn=9780520929074.

Herring, D. J., & Lynch, C. (2014). Law student learning gains produced by a writing assignment and instructor feedback. *Legal Writing: Journal of Legal Writing Inst, 19*, 103–126 Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jlwriins19&section=7.

Kosse, S. H., & ButleRitchie, D. T. (2003). How judges, practitioners, and legal writing teachers assess the writing skills of new law graduates: a comparative study. *Journal of Legal Education, 53*(1), 80–102 Retrieved from http://www.jstor.org/stable/42893788.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: an academic literacies approach. *Studies in Higher Education*, *23*(2), 157–172. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/03075079812331380364.

Lillis, T., & Turner, J. (2001). Student writing in higher education: contemporary confusion, traditional concerns. *Teaching in Higher Education*, *6*(1), 57–68. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13562510020029608.

Lisacek, F., Chichester, C., Kaplan, A., & Sandor, Á. (2005). Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM)* (pp. 41–50). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.477.6519&rep=rep1&type.pdf.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

Murumba, S. (1991). Good legal writing: a guide for the perplexed. *Monash UL Rev., 17*(1), 93–105 Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/monash17&section=11.

National Commission On Writing. (2003). *Report of the national commission on writing in America's schools and colleges: the neglected "R," the need for a writing Revolution*. College Board. Retrieved from http://www.collegeboard.com/prod_downloads/writingcom/neglectedr.pdf

Niedwiecki, A. (2006). Lawyers and learning: a metacognitive approach to legal education. *Widener Law Review, 13*(1), 33–73 Retrieved from http://widenerlawreview.org/files/2011/02/02NIEDWIECKI.pdf.

Niedwiecki, A. (2012). Teaching for lifelong learning: improving the metacognitive skills of law students through more effective formative assessment techniques. *Capital University Law Review, 40*(1), 149–194 Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2302717.

Norton, L. S. (1990). Essay-writing: what really counts? *Higher Education, 20*(4), 411–442 Retrieved from http://link.springer.com/article/10.1007/BF00136221.

OECD, & Statistics Canada (2010). *Literacy in the information age - final report of the international adult literacy survey*. OECD .Retrieved from http://www.oecd.org/edu/skills-beyond-school/41529765.pdf

Osbeck, M. K. (2012). What is "good legal writing"and why does it matter? *Drexel Law Review, 4*(2), 417–466 Retrieved from http://bits.rulebase.com/wp-content/uploads/2014/07/Good-Legal-Writing.pdf.

Parker, C. M. (1997). Writing throughout the curriculum: why law schools need it and how to achieve it. *Nebraska Law Review, 76*(3), 561–603 Retrieved from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1515&context=nlr.

QSR International Pty Ltd. (2012). NVivo qualitative data analysis software (Version 10) [Windows]. Retrieved from https://www.qsrinternational.com/products_nvivo.aspx.

Samuelson, P. (1984). Good legal writing: of Orwell and window panes. *University of Pittsburgh Law Review, 46*(1), 149–170 Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/upitt46&section=13.

Sándor, Á. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée, 12*(2), 97–108 Retrieved from http://www.cairn.info/load_pdf.php?ID_ARTICLE=RFLA_122_0097.

Sándor, Á., & Vorndran, A. (2009). Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries* (pp. 36–44). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1699757.

Sándor, Á., Kaplan, A., & Rondeau, G. (2006). Discourse and citation analysis with concept-matching. In *International Symposium: Discourse and document (ISDD)* (pp. 15–16). Retrieved from http://www.xrce.xerox.com/content/download/16625/118566/file/result.pdf.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: current applications and new directions*. New York: Routledge.

Shum, S. B., Sándor, Á., Goldsmith, R., Wang, X., Bass, R., & McWilliams, M. (2016). Reflecting on reflective writing analytics: Assessment challenges and iterative evaluation of a prototype tool. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 213–222). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2883955.

Simsek, D., Sandor, A., Shum, S. B., Ferguson, R., De Liddo, A., & Whitelock, D. (2015). Correlations between automated rhetorical analysis and tutors' grades on student essays. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 355–359). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2723603.

Sperling, C., & Shapcott, S. (2012). Fixing students' fixed mindsets: paving the way for meaningful assessment. *Legal Writing: Journal of Legal Writing Inst., 18*, 39–84 Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jlwriins18&section=6.

Stark, S. (1984). Why lawyers can't write. *Harvard Law Review, 97*(6), 1389–1393.

Sullivan, W. M., Colby, A., Welch Wegner, J., Bond, L., & Shulman, L. S. (2007). Educating lawyers: preparation for the profession of law. Stanford, California: Carnegie Foundation for the Advancement ofTeaching. Retrieved from http://archive.carnegiefoundation.org/pdfs/elibrary/elibrary_pdf_632.pdf.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Swales, J. M., Feak, C. B., Committee, S. C. D., Council, S, & others (2004). *Academic writing for graduate students: Essential tasks and skills (Vol*, *1*. MI: University of Michigan Press Ann Arbor Retrieved from http://www.tesl-ej.org/wordpress/issues/volume8/ej32/ej32r1/?wscr.

The Task Force on Law Schools and the Profession: Narrowing the Gap. (1992). *Legal education and professional development - an educational continuum*. American bar association. Retrieved from http://www.americanbar.org/content/dam/aba/publications/misc/legal_education/2013_legal_education_and_professional_development_maccrate_report%29.authcheckdam.pdf.

Todd, A. G. (2013). Writing lessons from abroad: a comparative perspective on the teaching of legal writing. *Washburn LJ, 53*(2), 295–326 Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/wasbur53&section=16.

Vinson, K. E. (2005). Improving legal writing: a life-long learning process and continuing professional challenge. *Touro Law Review, 21*(2), 507–550 Retrieved from http://papers.ssrn.com/abstract=847644.

Xerox. (n.d.). Xerox incremental parser. Retrieved October 29, 2015, from https://open.xerox.com/Services/XIPParser.